Qiong Hu (405065032)
Zihao Zou (005349580)
Gaofang Sun (104853165)

Project 2 Report

Social Network Mining

ECE 232E
Large Scale Networks
May 5, 2020

In this project, we will study the various properties of social networks. In the first part of the project, we will study an undirected social network (Facebook). In the second part of the project, we will study a directed social network (Google +).

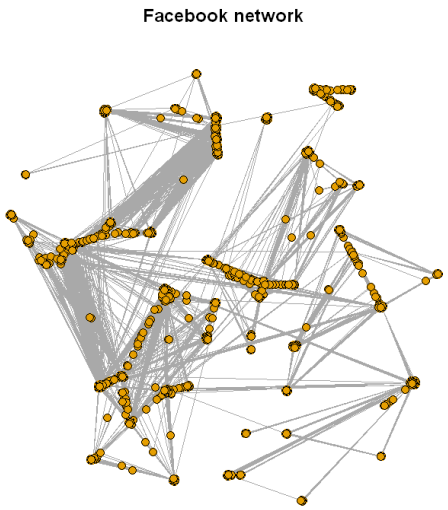
1. Facebook network

We obtained our Facebook dataset from http://snap.stanford.edu/data/egonets-Facebook.html and unzipped the edgelist file facebook_combined.txt.gz. we first create the Facebook network from edgelist file to learn the structural properties of the Facebook network. Then compile the personalized network and core node's personalized network and the friend recommendation in personalized networks.

1. Structural properties of the Facebook network

By creating the Facebook network, we will study the connectivity and degree distribution of the network.

QUESTION 1: Here is the graph for the Facebook network:



QUESTION 1.1:

The number of nodes is 4039.

The number of edges is 88234.

QUESTION 1.2:

The graph is connected.

QUESTION 2:

The diameter of the network is 8.

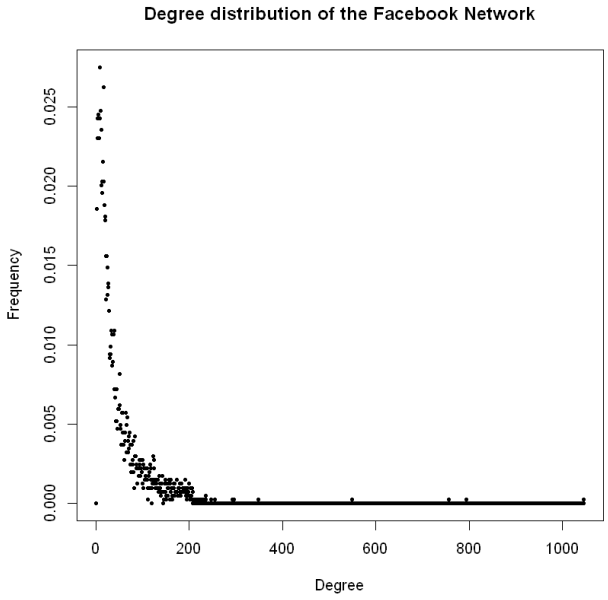
Qiong Hu (405065032)
Zihao Zou (005349580)
Gaofang Sun (104853165)

Project 2 Report

Social Network Mining

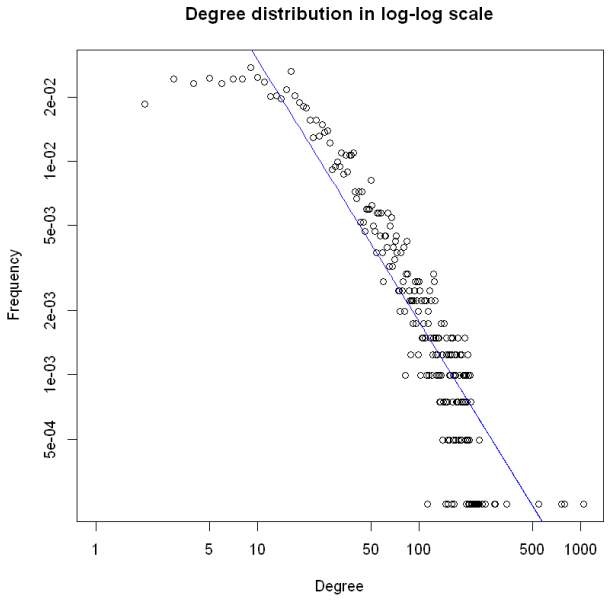
ECE 232E
Large Scale Networks
May 5, 2020

QUESTION 3: Here is the degree distribution of the Facebook network.



From the plot, we can see that most nodes have low degrees. As the node degree number increases, the frequency decreases. The average degree is 43.69.

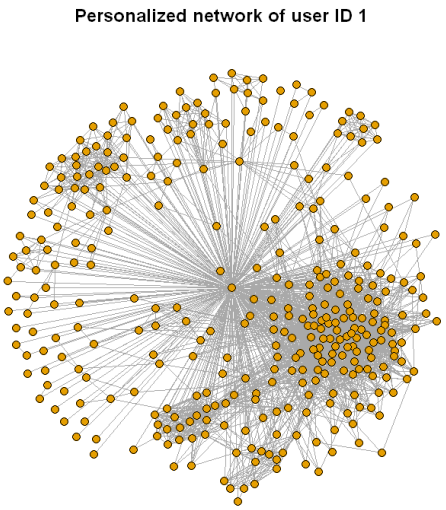
QUESTION 4: Here we plot the degree distribution of the Facebook network in log-log scale. We fit a line to estimate the slope of the trend of degree frequency. We did a linear fit in log-log scale and get the slope, -1.2279. The fitted line is in blue as shown below.



2. Personalized network

In this part, we will study some of the structural properties of the personalized network of the user whose graph node ID is 1 (node ID in edgelist is 0). From this point onwards, whenever we are referring to a node ID we mean the graph node ID which is 1 + node ID in edgelist.

QUESTION 5: Here is the personalized network of the user whose ID is 1.



The number of nodes is 348.

The number of edges is 2866.

QUESTION 6:

The diameter of this personalized network is 2. A trivial upper bound for the diameter of the personalized network is 2 and a trivial lower bound is 1.

QUESTION 7:

"Diameter" means the "longest shortest path" between any two vertices. In the context of the personalized network, the diameter to be equal to the upper bound "2" means that there are at least two nodes in the graph, between which the shortest path is 2. In other words, there exists at least a pair of two people that are not mutual friends, thus their shortest path goes through the center user of the personalized network, because both of them should be friends of the center user by the definition of "personalized network".

The diameter to be equal to the lower bound "1" means that the graph is fully connected and all nodes are connected with all other nodes, which also means all friends of the center user are friends with each other.

3. Core node's personalized network

In this part, we define a core node being the node that have more than 200 neighbors. We study various properties of the personalized network of the core nodes.

Qiong Hu (405065032)
Zihao Zou (005349580)
Gaofang Sun (104853165)

Project 2 Report

Social Network Mining

ECE 232E
Large Scale Networks
May 5, 2020

QUESTION 8:

The number of core nodes in the Facebook network is 40.

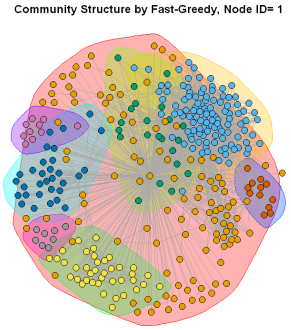
The average degree of those core nodes is 279.375.

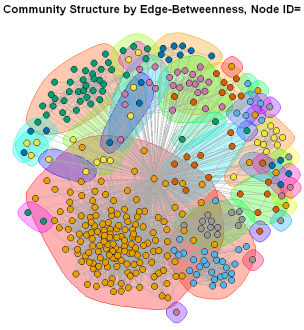
3.1. Community structure of core node's personalized network

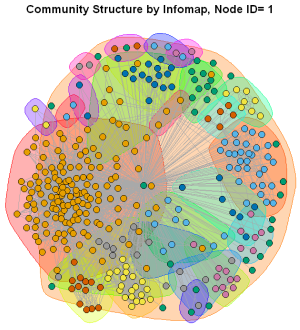
In this part, we study the community structure of the core node's personalized network. To be specific, we will study the community structure of the personalized network of the following core nodes:

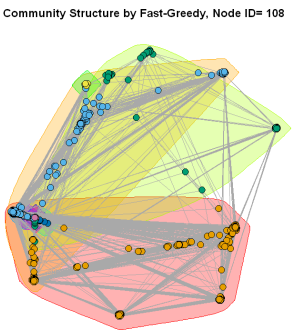
- Node ID 1
- Node ID 108
- Node ID 349
- Node ID 484
- Node ID 1087

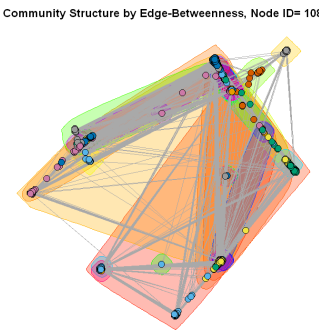
QUESTION 9: For each core node listed above, we used Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithms to find community structure. Here are 15 plots of the community structure of the core node's personalized networks:

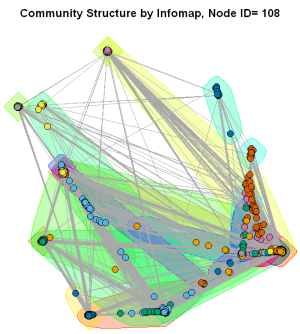


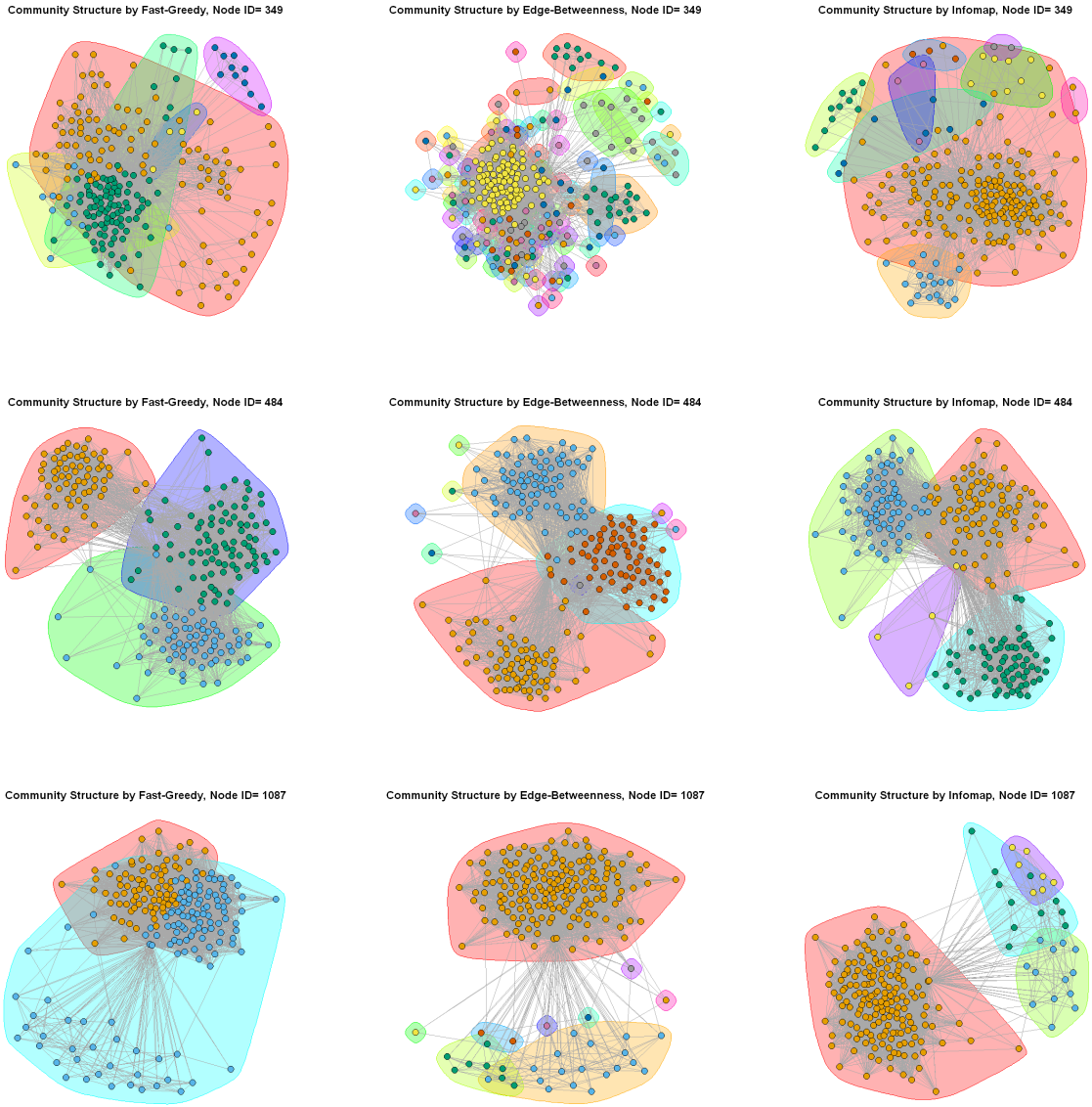












Core node ID	1	108	349	484	1087
Fast-Greedy	0.4131	0.4359	0.2517	0.5070	0.1455
Edge-Betweenness	0.3533	0.5068	0.1335	0.4891	0.0276
Infomap	0.3891	0.5082	0.0955	0.5153	0.0269

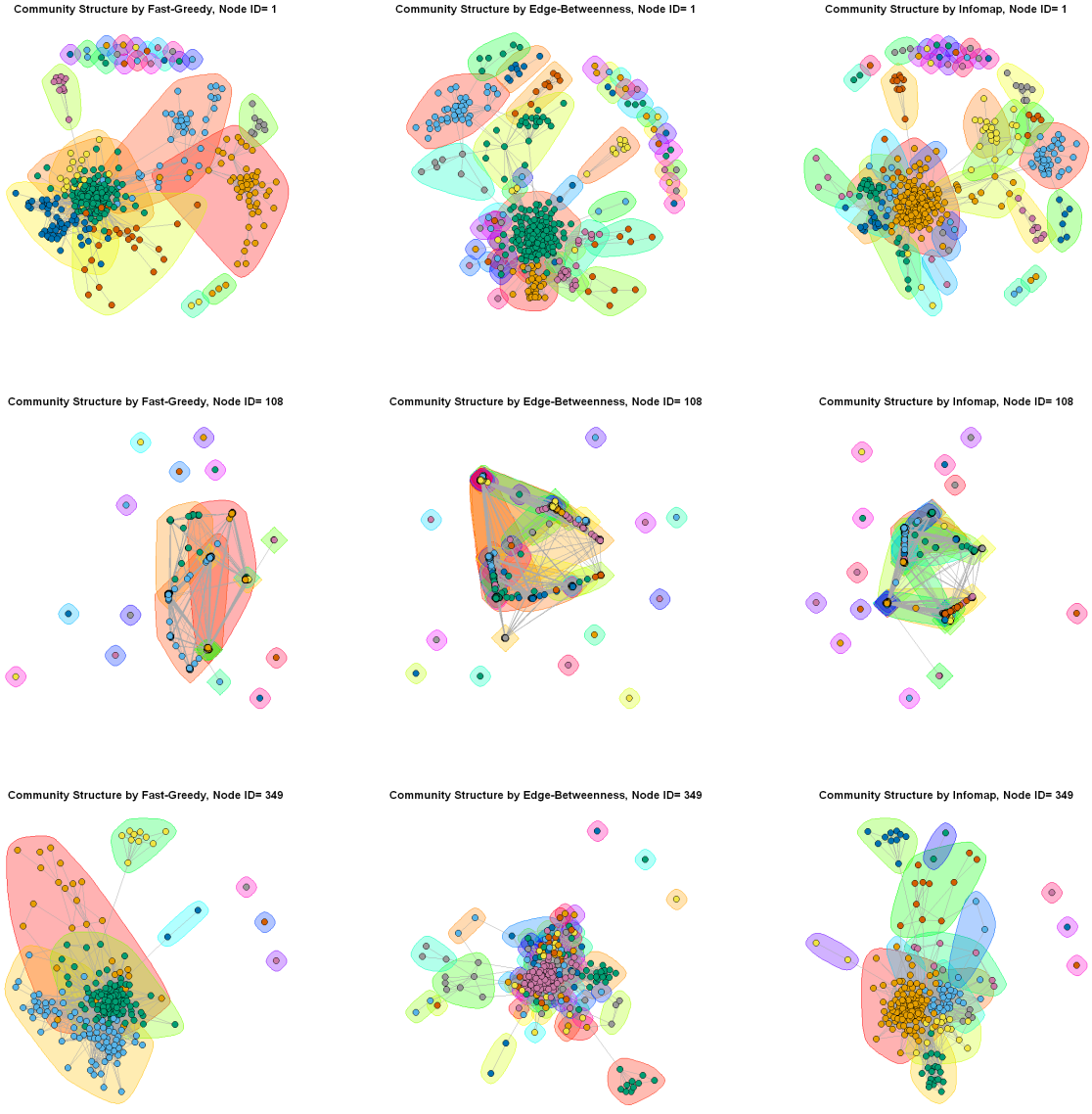
Table 1: The modularity scores of the algorithms

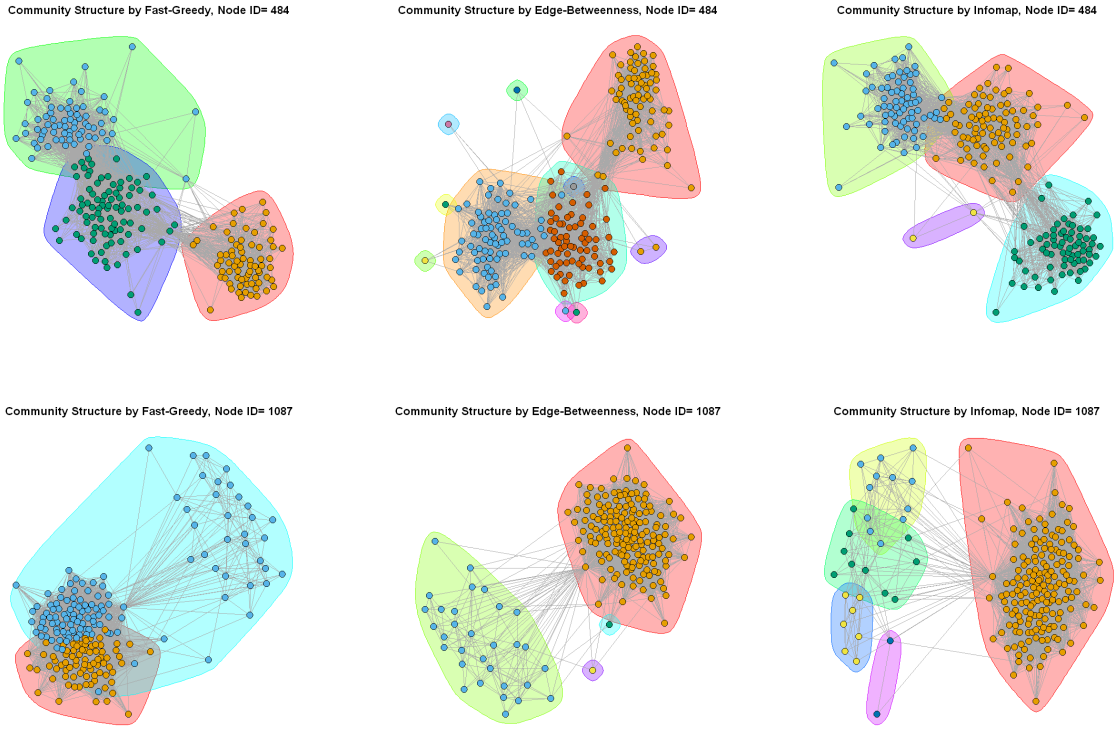
In most cases, the difference of the modularity scores between Edge-Betweenness and Infomap community detection algorithms are smaller than the difference between these two and Fast-Greedy algorithm.

3.2. Community structure with the core node removed

In this part, we explored the effect on the community structure of a core node's personalized network when the core node itself is removed from the personalized network and compare the modularity scores with Question 9.

QUESTION 10:





Core node ID	1		108		349	
Core node	w/	w/o	w/	w/o	w/	w/o
Fast-Greedy	0.4131	0.4418	0.4359	0.4581	0.2517	0.2457
Edge-Betweenness	0.3533	0.4161	0.5068	0.5213	0.1335	0.1506
Infomap	0.3891	0.4180	0.5082	0.5205	0.0955	0.2448

Core node ID	484		1087	
Core node	w/	w/o	w/	w/o
Fast-Greedy	0.5070	0.5342	0.1455	0.1482
Edge-Betweenness	0.4891	0.5154	0.0276	0.0325
Infomap	0.5153	0.5434	0.0269	0.0274

Table 2: The modularity scores of the algorithms with/without core nodes

Except for the core node ID=349 using Fast-Greedy algorithm, all the other core nodes and community detection algorithms report a larger modularity score when the core node is removed. The reason is that, in most cases, after the core node is removed, the personalized network becomes more scattered and is not necessarily connected, thus the communities have a more clear structure, resulting to a higher modularity score.

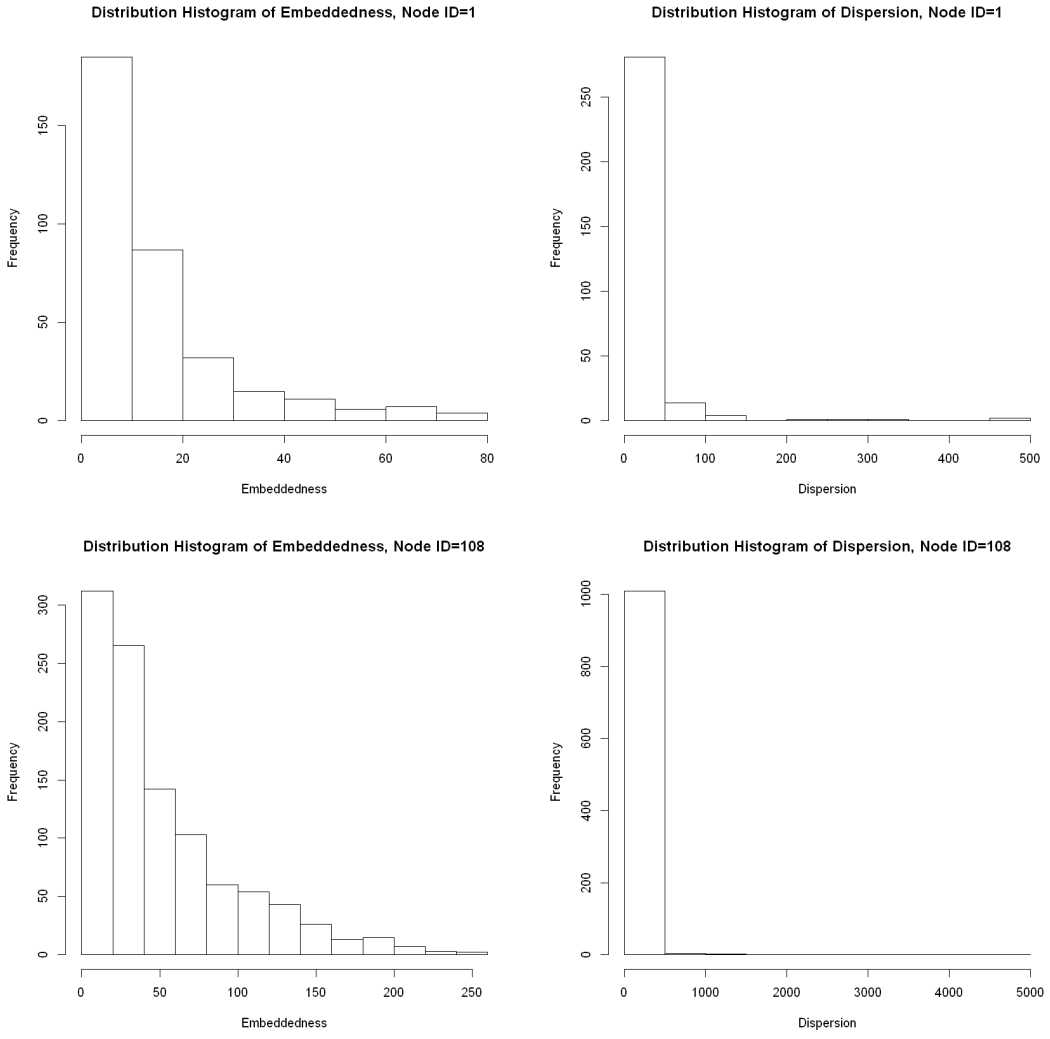
3.3. Characteristic of nodes in the personalized network

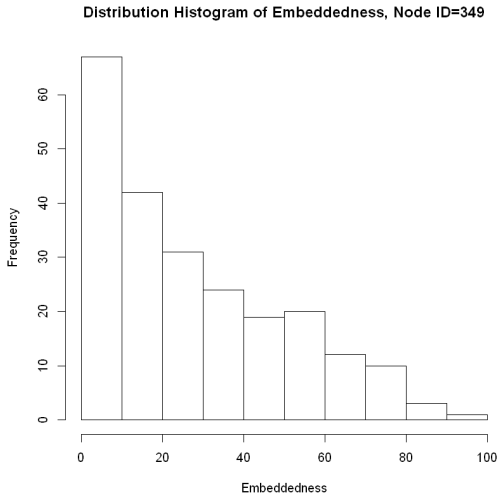
QUESTION 11: Because every neighbor of the non-core node in the core node's personalized network would either be a neighbor of the core node or the core node itself, due to the definition of the personalized network. Therefore, the expression between the non-core node's Embeddedness and its degree in the personalized network is that:

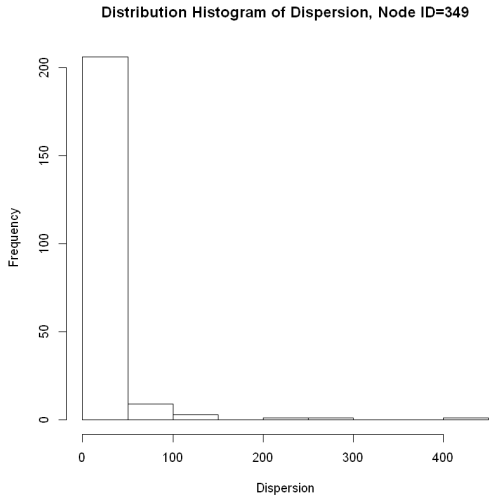
$$Embeddedness = Degree - 1. \tag{1}$$

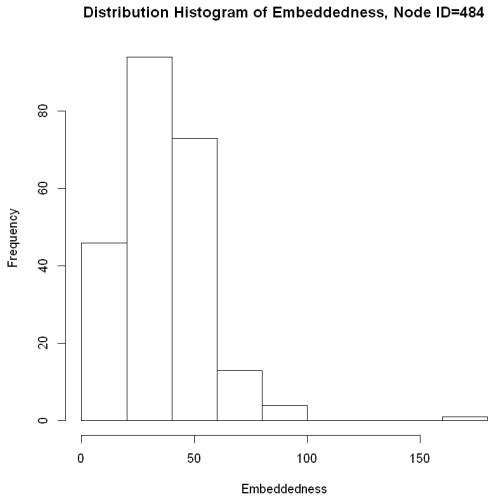
QUESTION 12: The distance in Dispersion is defined the same as [1]: The distance between a pair of mutual friends the non-core node shares with the core node equal to 1 when they are not directly linked and also have no common neighbors in the personalized network other than the core node and the non-core node (whose dispersion is being computed), and equal to 0 otherwise.

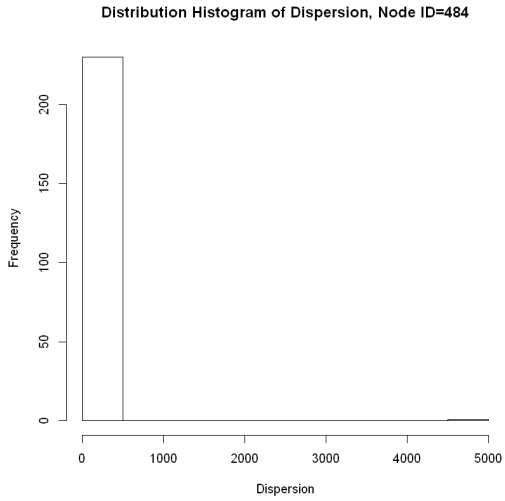
The distribution histogram of embeddedness and dispersion in the five personalized networks are shown below.

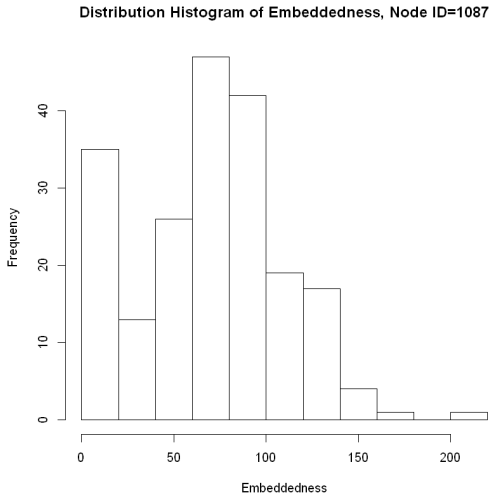


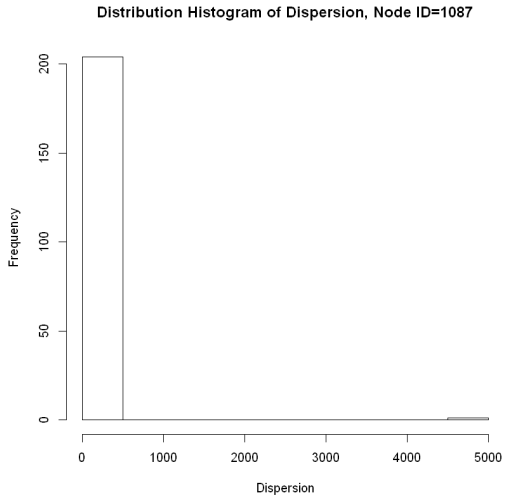

Distribution Histogram of Embeddedness, Node ID=349
Frequency
Embeddedness
0
10
20
30
40
50
60
0
20
40
60
80
100


Distribution Histogram of Dispersion, Node ID=349
Frequency
Dispersion
0
50
100
150
200
0
100
200
300
400


Distribution Histogram of Embeddedness, Node ID=484
Frequency
Embeddedness
0
20
40
60
80
0
50
100
150

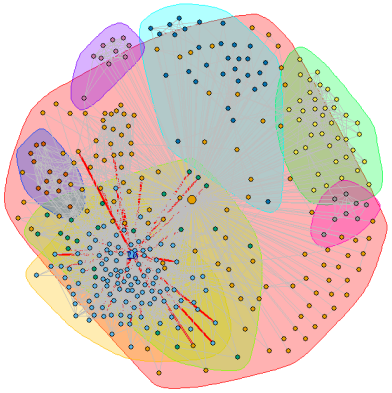

Distribution Histogram of Dispersion, Node ID=484
Frequency
Dispersion
0
50
100
150
200
0
1000
2000
3000
4000
5000


Distribution Histogram of Embeddedness, Node ID=1087
Frequency
Embeddedness
0
10
20
30
40
0
50
100
150
200

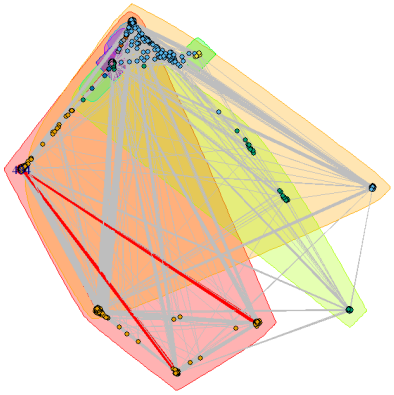

Distribution Histogram of Dispersion, Node ID=1087
Frequency
Dispersion
0
50
100
150
200
0
1000
2000
3000
4000
5000

QUESTION 13: The node with maximum dispersion and the core node are marked in a larger size, and the incident edges of that max dispersion node are highlighted in red thick lines.

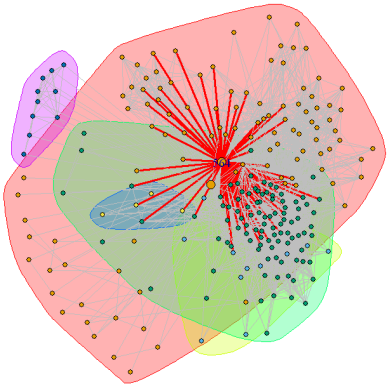
Community with highlighted max dispersion node, ID=1



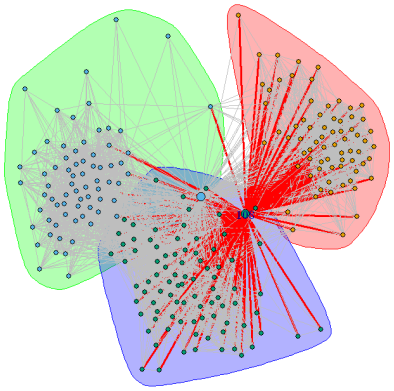
Community with highlighted max dispersion node, ID=108



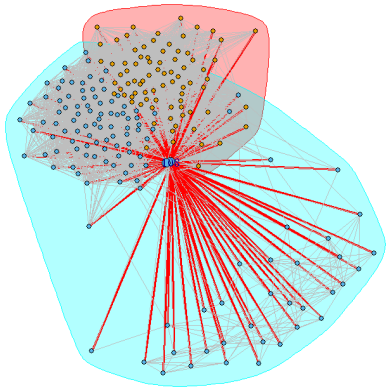
Community with highlighted max dispersion node, ID=349



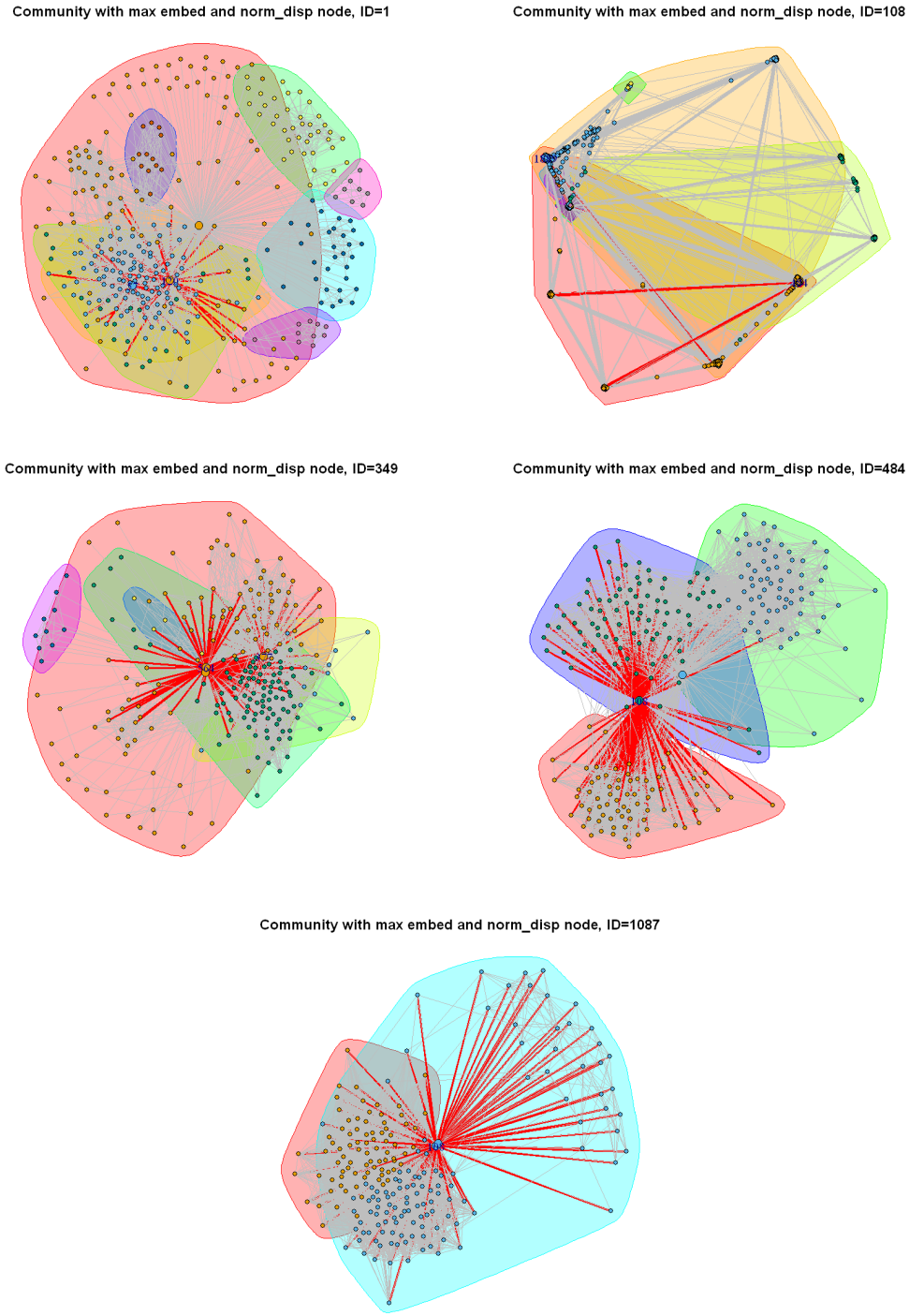
Community with highlighted max dispersion node, ID=484



Community with highlighted max dispersion node, ID=1087



QUESTION 14: The node with maximum embeddedness and normalized dispersion ($\frac{\text{dispersion}}{\text{embeddedness}}$) and the core node are marked in a larger size. The incident edges are highlighted in red thick lines. The node IDs are reported in Table 3 below.



Core node ID	1	108	349	484	1087
max embeddedness	57	1889	377	108	108
max dispersion	26	484	564	108	108
max normalized dispersion	120	484	564	108	108

Table 3: Maximum embeddedness, dispersion, normalized dispersion node ID

QUESTION 15:

Both embeddedness and dispersion can be used to characterize the tie strength between a node and the core node in its personalized network. In the application, they are used to describe the "closeness" of the relationships between these two nodes.

The embeddedness refers to the number of mutual friends a node shares with the core node. The larger the embeddedness is, the more mutual friends they share. In the application of facebook network, nodes in the same community means people in the same social foci, where people in the same cluster tend to know each other, and thus the large embeddedness doesn't necessarily represent particularly strong ties.

The dispersion reveals the importance of the pair of nodes to connect their mutual friends. In real life, people in relationship or partnership may not necessarily have maximum mutual friends, but their mutual friends tend to be from different social clusters, and the only connection between these clusters are those two people. Therefore, the larger the dispersion is, the stronger the tie between these two nodes are.

To combine the two measures together, we calculate the normalized dispersion ($\frac{\text{dispersion}}{\text{embeddedness}}$) since the absolute dispersion is normalized by the embeddedness. According to the previous analysis, the node with maximum normalized dispersion has the strongest tie to the core node. In our example of the five core nodes, we predict that the node 108 and 484 have a high probability to be in an intimate relationship, since they are each other's maximum normalized dispersion node.

4. Friend recommendation in personalized networks

4.3. Creating the list of users

QUESTION 16:

We created the list of users with degree 24, and denote the list as N_r.

$|N_r| = 11$.

4.4. Average accuracy of friend recommendation algorithm

QUESTION 17: In this question, three friend recommendation algorithms are implemented, the Common Neighbor, Jaccard, and Adamic Adar. For the Common Neighbors, the intersection of the lists of neighbors between two nodes are measured. For the Jaccard, the score was given by the intersection divided by the union of the lists of neighbors between two nodes. For the Adamic

Adar, the score is given by the sum of one over the log of the intersection of the lists of neighbors between two nodes. The following are the formal equations for the three methods. S represents the set of neighbors that node i or j has.

$$\text{Common Neighbors } (i,j) = |S_i \cap S_j| \tag{2}$$

$$\text{Jaccard}(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{3}$$

$$\text{Adamic Adar } (i,j) = \sum_{k \in S_i \cap S_j} \frac{1}{\log(|S_k|)} \tag{4}$$

To apply the friend recommendation algorithms, we randomly delete nodes and edges in the target graph with the probability of 0.25. Then with the friend recommendation algorithms, we recommend the same number of friends that got deleted to the target node. Afterward, we compare the deleted list of node and the recommended list of node to assess the accuracy of each algorithm. The following is the table of accuracy accordingly.

Node ID	Common Neighbors	Jaccard	Adamic Adar	Average
497	0.304	0.161	0.352	0.272
579	0.990	0.940	0.990	0.973
601	0.905	0.855	0.888	0.882
616	0.821	0.771	0.836	0.810
619	0.438	0.530	0.416	0.461
628	1.000	0.950	0.975	0.975
644	0.889	0.903	0.897	0.896
659	0.997	0.927	0.968	0.964
660	0.994	0.945	0.988	0.975
662	0.894	0.904	0.882	0.894
663	0.946	0.918	0.973	0.946
Average	**0.834**	**0.800**	**0.833**	0.823

Table 4: The accuracy for each algorithm and each node

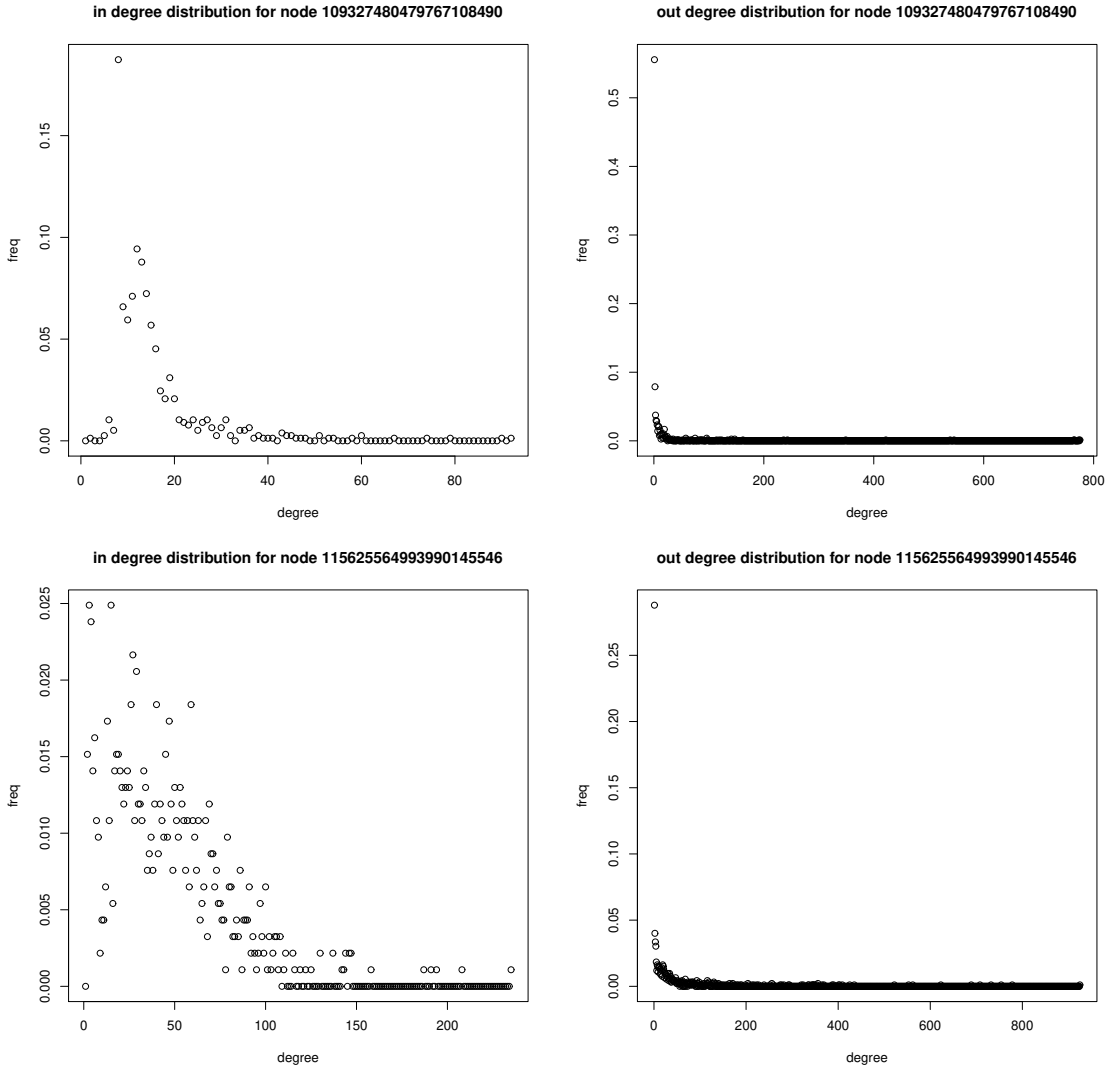
From the table above, we can see that all three of the algorithms have an average accuracy of over 80%. The common neighbors algorithm has a slightly better average accuracy than Adamic Adar and Jaccard, and occasionally can achieve 100% accuracy for certain nodes.

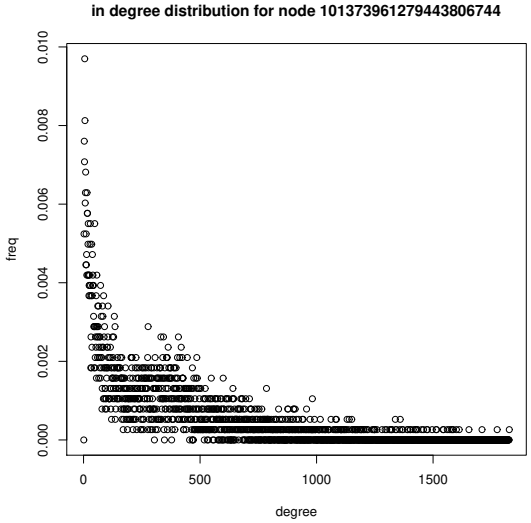
2. Google+ network

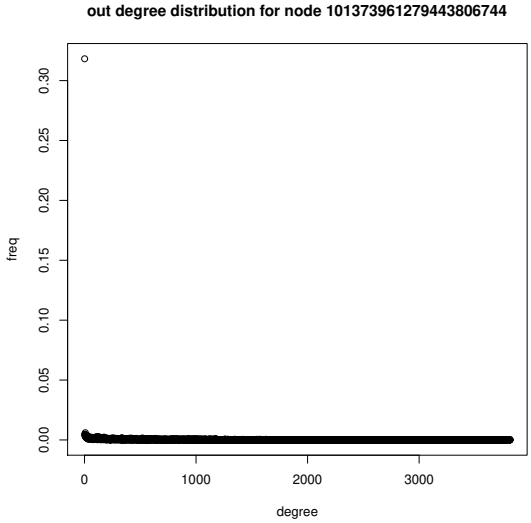
QUESTION 18: After creating the directed personal networks for users who have more than 2 circles, we have 57 personal networks in total.

QUESTION 19: To obtain the in and out degree of each graph, the ".edges" files are read with the "ncol" format. After reading the edges, the ego node is added. Stated in the data configuration, the ego node follows every other nodes. Therefore, an edge list of the ego node is alse created and added to the graph. The in-degree and out-degree distributions of three nodes with given ID are plotted below.

They have very different in-degree distributions, and a similar out-degree distribution.

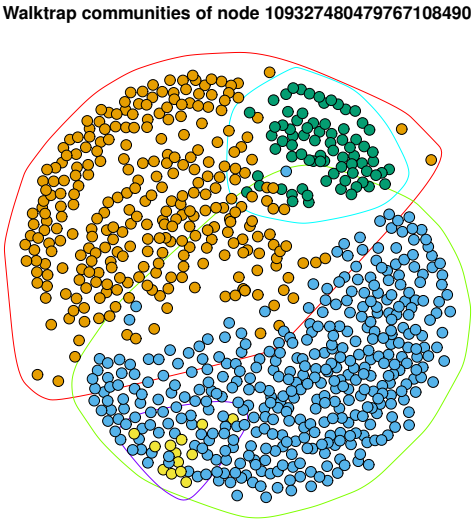


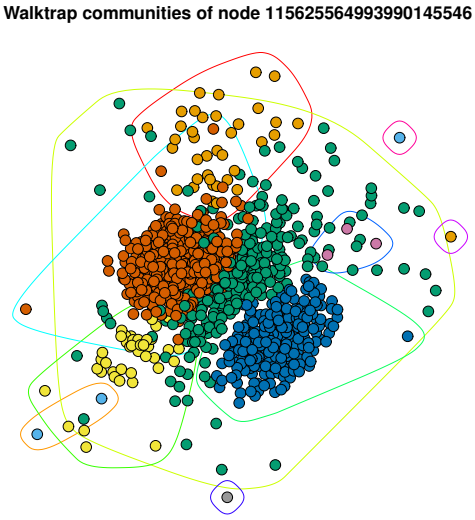


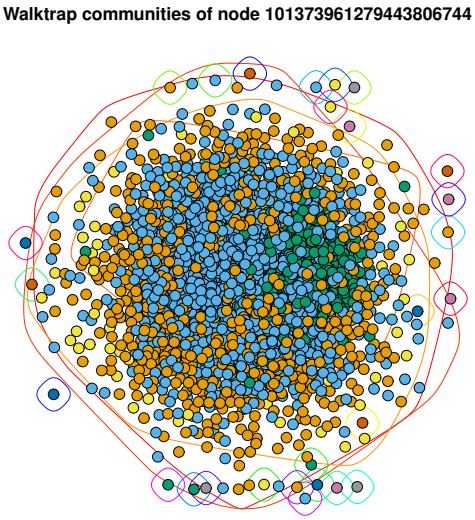


1. Community structure of personal networks

QUESTION 20: For each ego graph, the community structure is extracted with the Walktrap community detection algorithm and plotted below. Since the community structures can be complex, a transparent background is used in rendering. The modularity is reported in Table 5.







Node Id	Modularity
109327480479767108490	0.2527806
115625564993990145546	0.3194738
10137396127944380674	0.1910934

Table 5: The modularity of three personal networks

We found that, the more interconnected each node is in the community, the higher the modularity is. The first and second graphs have similar scores suggesting that nodes in each community is connected in a similar way. Moreover, we can tell that the third graph, although large, is actually less densely connected and results in a lower modularity score.

QUESTION 21: Homogeneity assesses if each community contains only members of a single circle. High homogeneity score shows that each community contains only members from a single circle. A low score means that each community contains members from different classes and knowing which community the user is from does not help determining which circles do they belong.

Completeness is almost the opposite to homogeneity. To receive a high completeness score, all members of a circle will have to be covered by the same community and each community can contain a subset of circles.

QUESTION 22: The following table shows the homogeneity and completeness scores accordingly.

Node ID	109327480479767108490	115625564993990145546	10137396127944380674
Homogeneity	0.5249	0.1060	0.0003744
Completeness	0.5497	0.3822	0.0008590

Table 6: The homogeneity and completeness scores for each node

For the node 109327480479767108490, both the homogeneity and completeness scores are relatively

high comparing to the rest of the nodes. This can mean that both the numbers of communities and circles are are relatively small. If the number of circles is small, the homogeneity score will go up since each community will only contain members from a few circles. Vice Versa, if the number of community is small, all members of the same circle are more likely to stay in one community. In contrast, the graph for node 101373961279443806744 is relatively large and contains more circles/communities. Hence, both the homogeneity and completeness scores are very small. Each community is large and can contain members from different circles. Similarly, the members of a circle are more likely to be spread around since there are more communities to choose from. For the middle node, 115625564993990145546, the homogeneity score is lower than the completeness score. A low homogeneity score means that each community is large and is able to cover members from different circles. Yet, the high completeness scores shows that the number of communities might not be high and members from each circle are not spread out to different communities.

References

[1] Backstrom, L. and Kleinberg, J. (2014, February). "Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook." In *Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing* (pp. 831-841).